CLAIMS

What is claimed is

1. A resource assignment method comprising:

establishing a resource model;

acquiring an application model; and

utilizing a mapping process to map said application model onto said resource model, wherein said mapping process is directed to increasing the optimization of resource utilization through appropriate assignment of resources to an application with respect to desired objectives.

2. The resource assignment method of Claim 1 further comprising

obtaining a set of parameters associated with topology and performance characteristics of resources in a data center and;

acquiring information about resource requirements of an application.

3. The resource assignment method of Claim 2 wherein said parameters that characterize the topology and resources of said data center include:

the number of edge switches, the number of rack switches, the number of server nodes, and connectivity matrices between different layers; and

specification of the bandwidth limits of the incoming and outgoing links at various layers of the network.

4. The resource assignment method of Claim 2 wherein said information about resource requirements of an application include:

the number of application functional components;

the network traffic requirements between said application functional components; and

upper and lower bounds on server attributes which are required for said server to host said application functional component.

5. The resource assignment method of Claim 1 wherein said mapping process determines which server nodes are assigned to an application functional component and is captured in an assignment decision variable.

6. The resource assignment method of Claim 5 wherein said assignment decision variable is optimized in accordance with a desired objective including meeting application requirements.

7. The resource assignment method of Claim 5 wherein said desired objective further includes minimizing communication delays.

8. The resource assignment method of Claim 5 wherein a layered partitioning and pruning (LPP) process is utilized to find an application resource assignment optimal solution.

9. A computer readable medium comprising instructions which when executed by a computer system causes the computer to implement an application resource mapping process comprising:

determining if there are enough feasible servers;

analyzing if a desirable assignment configuration is available if there are enough feasible servers;

saving an optimal feasible assignment variable in an application mapping template if a desirable assignment configuration is available;

sending said optimal feasible assignment variable to an application deployment service; and

computing remaining resources and updating a resource configuration template.

10. The computer readable medium comprising instructions which when executed by a computer system causes the computer to implement an application resource mapping process of Claim 9 further comprising

providing an indication there are not enough feasible servers if there are not enough feasible servers, and sending said application to another portion of a data center.

11. The computer readable medium comprising instructions which when executed by a computer system causes the computer to implement an application resource mapping process of Claim 9 wherein parameters are entered into a pruning algorithm to search for an optimal feasible assignment variable.

12. The computer readable medium comprising instructions which when executed by a computer system causes the computer to implement an application resource mapping process of Claim 9 further comprising providing an indication that there is not enough network bandwidth if there is not an optimal feasible assignment variable and sending said application to another portion of a data center.

13. The computer readable medium comprising instructions which when executed by a computer system causes the computer to implement an application resource mapping process of Claim 9 wherein the optimization problem is formulated with the objective of minimizing the average communication delay among the servers and the optimal

assignment also has to meet the constraints from the required server attributes and link bandwidth.

14. A computer readable medium comprising instructions which when executed by a computer system causes the computer to implement an application resource mapping process comprising:

determining if there is a need to add more servers;

performing a removal mapping process if no more servers are required;

performing an additional mapping process if more servers are required; and

deactivating a resource allocation service and waiting for a new request.

15. A computer readable medium comprising instructions which when executed by a computer system causes the computer to implement an application resource mapping process of Claim 14 wherein a current application mapping template is read and compared to the application requirements to obtain change requirements.

16. The computer readable medium comprising instructions which when executed by a computer system causes the computer to implement an application resource mapping process of Claim 14 further comprising:

calling a removal pruning algorithm to find an optimal set of servers to remove;

updating a application map file with a removal variable;

sending the application map file to a server removing service;

computing new remaining resources; and

updating a resource configuration file.

17. The computer readable medium comprising instructions which when executed by a computer system causes the computer to implement an application resource mapping process of Claim 14 further comprising:

determining if there are enough feasible servers;

calling an addition pruning algorithm to find an optimal set of servers to add an additional variable if there are enough feasible servers;

providing an indication that there is not enough network bandwidth if the search is not successful;

updating the application mapping file with the additional variable if the search is successful and sending the application mapping file to a server adding service;

computing remaining resources; and

updating the resource configuration file.

18. A resource allocation system comprising:

a means for gathering information associated with available networked resources;

a means for extracting information associated with application functional components; and

a means for assigning application functional components to said available networked resources in accordance with a resource allocation variable.

19. A resource allocation system of Claim 18 wherein said means for assigning application functional components to said available networked resources allocates said available networked resources by maximizing said available networked resources identified in said resource allocation variable with respect to application constraints and desired objectives.

20. The resource allocation system of claim 18 wherein said information associated with said available networked resources includes configuration and performance characteristics of said available networked resources.

21. The resource allocation system of claim 20 wherein said information associated with said application functional components includes the organization and networked resource requirements of said application functional components.

22. The resource allocation system of claim 21 wherein said means for assigning application functional components to said available networked resources includes a means for simplifying said assignment analysis by identifying infeasible networked resources and partitioning said available networked resources.